A Short Memo on Statistics

fwang2@ornl.gov

Last update: July, 2009

Contents

The summary is based on the following books I read:

- Head First Statistics, from Oreilly
- Probability and Statistics in Engineering by Hines, 3rd Edition.

1 Permutation and Combinations

If you choose r objects from a pool of n, the number of permutations is given by:

$$P(n,r) = \frac{n!}{(n-r)!} \tag{1}$$

If you choose r objects from a pool of n, the number of combination is given by:

$$C(n,r) = \frac{n!}{(n-r)!r!} \tag{2}$$

2 Summary of Data

Mean and Median

Mode The mode of a set of data is the most popular value, the value with the highest frequency. If there is more than one value has the highest frequency, then each such value is a mode.

If a data set has two modes, then we call the data "bimodal"

If you work for a small company, with a CEO and CTO, and bunch of small potatoes (workers), and you guys think your salary is too low and want to make an argument for the raise. Among all statistical averages (mean, median, and mode), which one should you use and why?

If CEO wants to make a counter argument, which one should he use?

Quartiles Quartiles are values that split your data into quarters (4 sections), so you have three such values. The lowest quartile is called the lower quartile, and the highest quartile is called the upper quartile. The

middle quartile is the median

Percentile Think of it as a generic case of quartile.

```
P25 - 25% percentile, Q1
P50 = 50% percentile, Q2
P75 = 75% precenttle, Q3
```

Variance a way of measuring spread, and it is the average of the distance of values from the mean sqaured.

$$var = \sum \frac{(x-\mu)^2}{n}$$

Standard deviation

$$\sigma = \sqrt{var}$$

Standard score, Standard error, z-score

$$z = \frac{x-\mu}{\sigma}$$

z can be seen as the number of standard deviation from the mean. Standard score work by transforming sets of data into a new, theoretical distribution with a mean of 0 and a standard deviation of 1. It is a generic distribution that can be used for comparing different data sets.

This score also provides a way of defining **outlier** - for example, if it is 3 standard deviation from the mean.

3 Probability

Definition

$$Probability = \frac{n(A)}{n(S)}$$

where $n(A)$ is the number for event A to happen and $n(S)$ is the sample space

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If event A and B are exclusive, $P(A \cap B) = 0$; If event A and B are exhaustive, $P(A \cup B) = 1$.

Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Visual Proof: think Venn Diagram

Law of Total Probability

$$P(B) = P(A) \times P(B|A) + P(A') \times P(B|A') \tag{3}$$

It gives you a way of finding the probability of a particular event based on on conditional probability.

Bayes Theorem

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A') \cdot P(B|A')} \tag{4}$$

4 Discrete Probability Distribution

Probability distribution it describes the probability of all possible outcomes of a given variable

Expectation Expected average long-term outcome

$$E(X) = \sum(x \cdot P(X = x)) \tag{5}$$

Variance

$$Var(X) = E(X - \mu)^2 = E(X^2) - \mu^2 \tag{6}$$

> If I want you to design a slot machine for a casino, it needs to serve two purpose: make sure you lose as much as you can; also entice you to keep playing ... in the context of probability, what do you need to do? hint: think $E(x)$ and Var ...

Linear transformation X and Y are old and new respectively.

$$E(aX + b) = aE(x) + b \tag{7}$$
$$Var(aX + b) = a^2 Var(X) \tag{8}$$
$$E(aX + bY) = aE(X) + bE(Y) \tag{9}$$
$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) \tag{10}$$
$$E(aX - bY) = aE(x) - bE(Y) \tag{11}$$
$$Var(aX - bY) = a^2 Var(X) + b^2 Var(Y) \tag{12}$$

Note that variance is addition even we are subtracting random variable, which is less intuitive.

Independent observations Think X is the gain variable for a slot machine, and you play multiple times, every time the outcome is an observation

$$E(X1 + X2... + Xn) = nE(X) \tag{13}$$
$$Var(X1 + X2... + Xn) = nVar(X) \tag{14}$$

5 A Few Important Discrete Distributions

5.1 Geometric distribution

When to use?

- you run a series of independent trials, either success or failure
- probability of success for each trial the same
- you are interested to how many trials needed to get your first success

Define: X is **number of trials needed to get the first success**. p is the probability of success in a trial,

$$X \sim Geo(p)$$

$$P(X = r) = pq^{r-1} \tag{15}$$
$$P(X > r) = q^r \tag{16}$$
$$P(X \leq r) = 1 - q^r \tag{17}$$
$$E(X) = 1/p \tag{18}$$
$$Var(X) = q/p^2 \tag{19}$$

The geometric distribution has a distintive shape: $P(X = r)$ is at its highest when $r = 1$, and gets lower and lower as r increases. Thus, the mode of any geometric distribution is always 1. This may be a bit counter-intuitive, as it says, it is mostly likely only one attempt will be needed for a successful outcome.

5.2 Binomial distribution

When to use?

- you run a series of independent trials, either success or failure
- probability of success for each trial the same
- you are interested in the number of successes in the n independent trials

Define: X is the number of successful outcomes out of n trials p is the probability of success in a trial,

$$X \sim B(n, p)$$

$$P(X = r) = C(n, r)p^r q^{(n-r)} \tag{20}$$
$$E(X) = np \tag{21}$$
$$Var(X) = npq \tag{22}$$

5.3 Poisson distribution

When to use?

- individual event occurs at random and independent at a given interval
- you know the mean number of occurrences in the interval (rate), and it is finite

- and you want to know the probability of certain number of occurrences in the interval

Define: X is the number of occurrences in a particular interval; λ is the rate of occurrences.

$$X \sim Po(\lambda)$$

$$P(X = r) = \frac{e^{-\lambda}(\lambda)^r}{r!} \tag{23}$$

$$E(X) = Var(X) = \lambda \tag{24}$$

If $X \sim Po(\lambda_x)$, $Y \sim Po(\lambda_y)$, X and Y are independent, then

$$X + Y \sim Po(\lambda_x + \lambda_y)$$

The shape of the Poisson distribution varies on λ. If λ is small, the distribution skews to the right, but it becomes more symmetric as λ increases. If λ is an integer, then there are two mode: λ and $\lambda + 1$. If λ is not an integer, then the mode is λ

5.4 Relationship between Poisson and Binomial

If $X \sim B(np)$ where n is large (> 50) and p is small (< 0.05) you can approximate it with $X \sim Po(np)$. Noted that np and npq are close the each other if q is close to 1 and n is large.

Why approximate? as calculating Binomial with large n is not as easy as with Poisson.

6 Continuous Variable and Distributions

6.1 Exponential Distribution

7 Normal Distribution

Also known as Gussian distribution. As this is the first continous distirbution, we want to highlight: for discrete probability distribution, we look at the probability of getting a particular **value**; for continuous probability distributios, we look at the probability of getting a particular **range**. For continous random variable:

$$\text{probability} = \text{area}$$

So to find $P(a < X < b)$, you need to calculate the area under the probability density function between a and b.

The normal distirbution is defined by two parameters: μ and σ^2. μ tells wher ethe center of the curve is, σ^2 gives you the spread. The larger σ^2 is, the flatter the shape is.

If random variable X follows normal distribution, it is often denoted as:

$$X \sim N(\mu, \sigma^2) \tag{25}$$

How to Calculate Probability for Normal Distribution

1. Determine μ and σ^2
2. Standarize to $N(0,1)$, which gives you a standarzied normal variable, Z, and $Z \sim N(0,1)$.
   The **standard score** of X is defined as:
   $$Z = \frac{X - \mu}{\sigma} \tag{26}$$
   When X takes on a specific value x, Z will take on an corresponding value of z, known as standard score of value x.
3. You can look up the usual probability table and the value listed there are for:
   $$P(Z < z)$$

As an example: $N(10,4), x = 6$. Following above equation, we have $z = (6-10)/\sqrt{4} = -2$. So -2 is the standard score of 6. The table lookup says $P(Z < -2) = 0.0228$.

Auxiliary functions are:

$$P(Z > z) = 1 - P(Z < z)$$
$$P(a < X < b) = P(X < b) - P(X < a)$$

Two normal variables:

If X $N(\mu_x, \sigma_x^2)$ and Y $N(\mu_y, \sigma_y^2)$, and X and Y are independent, then

$$X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$
$$X - Y \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$$

Notice that $X + Y$ and $X - Y$ have the same shape, except the mean is different.

Linear transform:

if $X \sim N(\mu, \sigma^2)$, a and b are numbers, then:

$$aX + b\, N(a\mu + b, a^2\mu^2)$$

Independent observations:

if $X_1, X_2, \ldots X_n$ are independent observations of X where

$$X \sim N(\mu, \sigma^2)$$

$$X1 + X2 + \ldots + Xn \; N(n * mu, n * sigma^2)$$

Noted that linear transformation affect the underlying value in your probability distribution; while independent observations have to do with the quantity of things you are dealing with.

7.1 Approximate Binomial Distribution with Normal

If $X \sim B(n, p)$ and $np > 5$, and $nq > 5$, you can use $X \sim N(np, npq)$ to approximate it.

Example: you want to take an exam with 100 questions, each question has 4 choices. If you just randomly select one answer out of 4. How do you calculate the probability that you can get half of the answers correct?

Continuity Correction

If you are approximating the binomial distribution with the normal distribution, then you need to apply a "continuity correction" to make your result accurate.

- if you work with $\geq$ or $\leq$, you need to make sure you include the value in the inequality in your probability range.
- If you work with $>$ or $<$, you need to make sure you exclude the value in the inequality from your probability range.
- So:

$$P(X < a) \rightarrow P(X < a - 0.5) \tag{27}$$
$$P(X = 0) \rightarrow P(-0.5 < X < 0.5) \tag{28}$$
$$P(X \geq a) \rightarrow P(X > a - 0.5) \tag{29}$$
$$P(X \leq a) \rightarrow P(X < a + 0.5) \tag{30}$$

7.2 Approximate Poisson Distribution

If $X \sim Po(\lambda)$ and $lambda > 15$, you can use $X \sim N(\lambda, \lambda)$ to approximate it. Remember: in this case, using continuous variable to approximate discrete variable, you need to apply continuity correction.

Example: The average number of downtime is 40 times per year for Jaguar machine. What is the probability of Jaguar is down less than 30 times a year?

- You can do the same as binomial, and apply continuity correction as above.
- We have done approximation for both binomial and Poisson using Normal. That's because if the certain conditions are met, the shape of Binomial and Poisson come very close to normal, but Geometric distribution never look like Normal, so forget about using Normal to approximate Geometric distribution.

7.3 Empirical Rules on Normal Distribution

- About 68% of your values lie within 1 standard deviation of the mean.
- About 95% of your values lie within 2 standard deviation of the mean.
- About 99.7% of your value lie within 3 standard deviation of the mean.

7.4 Chebyshev's Inequality

The Chebyshev inequality is presented as a bounding probability that a random variable lies between $\mu - k\sigma$ and $\mu + \sigma$. Its usefulness stems from the fact that so little knowledge is required on the random variable.

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2} \tag{31}$$

or

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

As an example: an inventory manager knows that the *lead time* required in ordering a part are 8 days and 1.5 days for mean and standard deviation. What would be the time interval during which there are over 90% of probability that he will receive the part? According to above equation: $1 - 1/k^2 = 0.9$. You would know k, then you know the interval $\mu \pm k\sigma$.

Of course, if you calculate with $k = 2, 3, 4$, then you have some rough idea on how data are distributed no matter what distribution they maybe: **at least** 75%, 89%, and 94% of your values lie within 2, 3, 4 standard deviation of the mean, respectively.

The problem with Chebyshev inequality is the kind of statement you can make is on the coarse side.

8 Sampling and Estimation

8.1 Sampling

Population A statistical population refers to the entire group of things that you are trying to measure, study.

Census It is a study or survey involving the entire population.

Sample A statistical sample is a selection of items taken from a population.

Unbiased sample is a representative of the target population. This means it has similar characteristic to the population, and we can use these to make inferences about the population itself.

Biased sample is not representative of the target population.

How to Design a Sample

1. Define your target population. So you know where you are collecting your sample from.
2. Define your sampling unit.
3. Define your sampling frame. A list of all sampling units within your target population, preferably with each sampling unit either named or numbered.

Sampling Method

1. **Simple random sampling**, with or without replacement.

2. **Stratified sampling**, the population is split into similar groups that share similar characteristics. These groups are called **strata**, and each individual group is called a **stratum**. Once you've done this, you perform simple random sampling **within** each stratum.

3. **Cluster sampling**, the population has a number of similar groups or clusters. Think a pack of cigarettes as the sampling unit.

4. **Systematic sampling**, you list population in some sort of order, and then survey every *i*-th item.

8.2 Estimation

Estimator for Mean

A **point estimator** of a population parameter is some function or calculation that can be used to estimate the value of the population parameter, derived from sample data. As an example, a point estimator of the population mean ($\hat{\mu}$ of μ) is the mean of the sample ($\bar{x}$).

Note that we differentiate between an actual population parameter and its point estimator using ˆ symbol. In this case, we use symbol μ to represent the population mean, and $\hat{\mu}$ to represent its estimator. When we estimate the population mean using the mean the sample, this means that:

$$\hat{\mu} = \bar{x} = \frac{\sum x}{n}$$

Estimator for Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

and

$$\hat{\sigma^2} = s^2$$

The formula for population variable point estimator is usually written as s^2. Diving a set of numbers by $n-1$ gives better result than dividing it by n, and this difference is more noticeable when n is small (having less samples).

Sample Distribution of Mean

Assume that we have been told the mean and variance of the gumball population in a pack, μ and σ^2. The number of gumballs in a pack is represented by X. Then each pack chosen at random is an **independent observation** of X. Let us take a sample of n gumball packs, X_1 through X_n, each has an expectation of μ and variance of σ^2.

If we know the sample mean distribution, we can answer question such as "what is the probability of having a mean of 8.5 gumballs or fewer in a sample of 30 packs of gumballs?".

$\overline{X}$ is the mean number of gumballs in each packets of gumballs in the sample, so:

$$\overline{X} = \frac{X_1 + X_2 + \ldots X_n}{n}$$

where each X_i represents number of gumballs in i packet of gumballs.

$$\begin{aligned} E(\overline{X}) &= E\big(\frac{X_1 + X_2 + \ldots + X_n}{n}\big) \\ &= E(\frac{1}{n}X_1) + E(\frac{1}{n}X_2) + \ldots + E(\frac{1}{n}X_n) \\ &= \frac{E(X_1) + E(X_2) + \ldots + E(X_n)}{n} \\ &= \frac{n\mu}{n} = \mu \end{aligned} \tag{32}$$

Similarly, variance can be calculated as:

$$Var(\overline{X}) = \frac{\sigma^2}{n} \tag{33}$$

Now comes to the key revelation: when n is large, $\overline{X}$ can still be approximated by the normal distribution.

$$X \sim N(\mu, \frac{\sigma^2}{n})$$

Estimator for Proportion

Say, you want to find out the proportion of a population (p) that likes your product. You conduct a survey on 40 people, 32 of them like it (or 32 success). If we use p_s to represent the proportion of successes in the sample, then we can estimate the proportion of successes in the population using:

$$\hat{p} = p_s = \frac{\text{number of successes}}{\text{number in sample}}$$

Sample Distribution of Proportion

Let's say we have the gumball population, and we have been told that the proportion of red gumball is $p = 0.25$.

Further, these gumballs are sold in packs, with each pack contains 100 gumballs. So, the sample size is 100, dented by n.

If we use random variable X to represent the number of red gumballs in the sample, then $X \sim B(n,p)$, where $n = 100$, and $p = 0.25$.

The proportion of red gumballs in the sample depends on X, the number of red gumballs in the samle. This means that **the proportion itself is a random variable**. If we write it as P_s, then $P_s = X/n$.

There are many possible sample we could take of size n. The number of red gumballs in each such sample is distributed as $B(n,p)$, and proportion of succcess is given by X/n.

We can form a distribution of all sample proportion using all possible samples. This is called **sampling distribution of proportions**, or P_s.

With this sampling distribution, we could answer questions such as "what is the probability that the proportion of red gumballs in one particular pack will be at least 40%"?

Expectation and Variance of P_s

Intuitively, expected value of sampling proportion should be match the proportion (of red gumballs) in the population, that is p. Let us check:

$$\begin{aligned} E(P_s) &= E(X/n) \\ &= \frac{E(X)}{n} \\ &= \frac{np}{n} \quad \text{since } X \sim B(n,p), \text{and } E(X) = np \\ &= p \end{aligned} \tag{34}$$

For variance, we follow similar reasoning:

$$\begin{aligned} Var(P_s) &= Var(X/n) \\ &= \frac{Var(X)}{n^2} \\ &= \frac{npq}{n^2} \quad \text{same as above }, Var(X) = npq \\ &= \frac{pq}{n} \end{aligned} \tag{35}$$

What is the Distribution for P_s?

The distribution of P_s depends on the size of the samples. Here is the highlight: when n is large (> 30), the distribution of P_s becomes approximately normal. Since we already deduced expectation and variance:

$$P_s \sim N(p, \frac{pq}{n})$$

You should do a continuity correction ($\pm 1/2n$, given that $P_s = X/n$ and continuity correction for X is $\pm 1/2$) when you use normal distribution to find probabilities. However, when n is large, this can be left out.

8.3 Introducing Central Limit Theorem

The central limit theorem says that if you take a sample from a non-normal population X, and if the size of the sample is large (> 30), then the distribution of $\overline{X}$ is approximately normal.

$$\overline{X} \sim N(\mu, \sigma^2/n) \tag{36}$$

The Binomial distribution

If we have a population represented by $X \sim B(n, p)$, where n is greater than 30. Then by CLT, we get:

$$\overline{X} \sim N(np, pq)$$

The Poisson distribution

If we have a population represented by $X \sim Po(\lambda)$, again where $n > 30$. Then by CLT, we get:

$$\overline{X} \sim N(\lambda, \lambda/n)$$

9 Confidence Interval

Besides point estimator, **confidence interval** is another way of estimating population statistics which allows for uncertainty. Confidence interval can be used on both continuous variable or categorical variable. If used on continuous variable, *mean* is the most often population estimate; on categorical variable, *proportion* is often the population estimate. For example, you can estimate the proportion of a certain outcomes given the sample population.

Here are the steps of finding it:

1. Choose population statistic, for which you construct confidence interval.
2. Finding its sampling distribution
3. Decide on the level of confidence, i.e., the probability your interval contains the statistic.
4. Find the confidence limit

Confidence interval is usually done for mean or population proportion. For example: $P(a < \mu < b) = 0.95$: there is 95% chance of the interval (a, b) containing the population mean.

Example: A sample size of 100, point estimation for mean is 62.7, variance is 25. We want to construct 95% confidence interval for mean.

1. Choose population statistic: that would be mean.

2. Find its sampling distribution: We know $E(X) = \mu$, and we know:

$$Var(\overline{X}) = s^2/n = 25/100 = 0.25$$

By CTL, sample mean distribution can be approxmiated by:

$$\overline{X} \sim N(\mu, 0.25)$$

3. Decide on the level of confidence: the probability of the population mean being inside the confidence interval is 95%.

4. Finding the confidence limits: think the bell curve, with μ in the center, a and b on the left and right, the enclosing area or the probability is 0.95. The tails on both end account for the rest of 0.05. This means $P(\overline{X} < a) = 0.025$ and $P(\overline{X} > b) = 0.025$.

We will find this interval by standarizing $\overline{X}$. We know

$$Z = \frac{\overline{X} - \mu}{\sqrt{0.25}} = \frac{\overline{X} - \mu}{0.5}$$

where $Z \sim N(0,1)$. We need to find z_a and z_b where $P(z_a < Z < z_b) = 0.95$.

- Using probability table to find value of z_a where $P(Z < z_a) = 0.025$, that gives us $z_a = -1.96$.
- Using probability table to find value of z_b where $P(Z > z_b) = 0.025$, that gives us $z_b = 1.96$.

Now we know that:

$$P\big(-1.96 < \frac{\overline{X} - \mu}{0.5} < 1.96\big) = 0.95$$

A few simple manipulation gives us:

$$P\big(\overline{X} - 0.98 < \mu < \overline{X} + 0.98\big) = 0.95$$

As $\overline{X}$ is the distribution of sample means, so we can use the value of $\bar{x}$ from the sample, which is 62.7. There you have it.

9.1 Shortcuts for Confidence Interval

In general, the confidence interval is given by:

$$\text{statistics} \pm (\text{margin of error})$$

Population statistics	Population distribution	Conditions	Confidence Interval
μ	Normal	You know what σ^2 is; n is large or small, $\bar{x}$ is the sample mean	$\left(\bar{x}-c\frac{\sigma}{\sqrt{n}}, \quad \bar{x}+c\frac{\sigma}{\sqrt{n}}\right)$
μ	Non-normal	You know what σ^2 is; n is large (at least 30), $\bar{x}$ is the sample mean	$\left(\bar{x}-c\frac{\sigma}{\sqrt{n}}, \quad \bar{x}+c\frac{\sigma}{\sqrt{n}}\right)$
μ	Normal or Non-normal	You don't know what σ^2 is; n is large (at least 30), $\bar{x}$ is the sample mean; s^2 is the sample variance	$\left(\bar{x}-c\frac{s}{\sqrt{n}}, \quad \bar{x}+c\frac{s}{\sqrt{n}}\right)$
p	Binomial	n is large; p_s is the sample proportion, q_s is $1-p_s$	$\left(p_s-c\sqrt{\frac{p_s q_s}{n}}, \quad p_s+c\sqrt{\frac{p_s q_s}{n}}\right)$
μ	Normal or non-normal	You don't know what σ^2 is; n is small (less than 30), $\bar{x}$ is the sample mean, s^2 is the sample variance	$\left(\bar{x}-t(\nu)\frac{s}{\sqrt{n}}, \quad \bar{x}+t(\nu)\frac{s}{\sqrt{n}}\right)$

The value of c depends on the level of confidence you need: for 90% confidence, $c = 1.64$; 95% confidence, $c = 1.96$, 99% confidence, $c = 2.58$.

9.2 Confidence Interval for Difference Between Two Means, and Differences Between Two Proportions

.

Both results are useful, to refresh if you actually refreshed, I omitted the deduction for your excises. You can assume population is normal distributions. The answer can be checked from head first book P652 and P653.

9.3 When Sample Size Is Small and σ^2 is Unknown...

Normal distribution is no longer a good approximation. **When population is normal, sample size is small, σ^2 is unknown,** , $\overline{X}$ follows the $t-$ distribution.

The $t-$ distribution looks like a smooth, symmetrical curve and its exact shape depend on the size of the sample: when sample size is large, it looks like a normal distribution, but when sample size is small, the curve is flatter and has slight fatter tails. Thus, when you use $t-$ distribution for estimating confidence intervals, you generally have a larger interval as situation is more uncertain.

It takes one parameter ν, where $\nu = n-1$. n is the size of the sample, ν is called the number of degree of freedom. A shorthand way of saying that T follows $t-$ distribution with ν degree of freedom is:

$$T \sim t(\nu)$$

You find the confidence limits with t-distirbution in a similar way as you do for normal distribution:

$$\left(\bar{x} - t\frac{s}{\sqrt{n}}, \quad \bar{x} + t\frac{s}{\sqrt{n}}\right),$$

where $P(-t \leq T \leq t) = 0.95$.

t-distribution probability table

t-distribution probability table give you the value of t where

$$P(T > t) = p$$

For 95% confidence level, $p = 0.025$.

10 Hypothesis Test

Hypothesis tests give you a way of using samples to test whether or not statistical claims are likely to be true. It involves the following steps:

1. Decide on the hypothesis you want to test
2. Choose the test statistic
3. Determine the critical region for your decision
4. Find the p-value of the test statistic
5. See wheter the sample result is within critical region
6. Make the decision: accept or reject the hyposis.

10.1 Walk through an example

An example: a drug company claims one of its wonder drug cures 90% of the people, a doctor collected data on 15 of his patient taking the drug: 11 cured, 4 is not. Should he accept the drug company's claim?

Step 1: Decide on the hypothesis

The claim we are testing is called **null hypothesis**, denoted by H_0. It is the claim we will accept unless there is strong evidence against it.

The counterclaim to the null hypothesis is called **alternative hypothesis**, denoted by H_1, and it is the claim we will accept if there is strong evidence to reject H_0.

In our case, the doctor wants to test:

$$H_0 : p = 0.9 \qquad H_1 : p < 0.9$$

Step 2: Choose test statistic

We need to decide what to test by assuming H_0 is true. The test statistic is the statistic that is most relevant to the test.

For this example, we can look at the whether the number of success in the sample is significant based on some probability distribution. If we use X to represent number of people cured in the sample, this means X is our test statisic. Each people in the sample is a independent test, with 0.9 probability of success. Then X follows a **binomial distribution**:

$$X \sim B(15, 0.9)$$

Step 3: Determine critical region

The **critical region** of a hypothesis test is the set of values that presetn the most extreme evidence **against the null hyposis**.

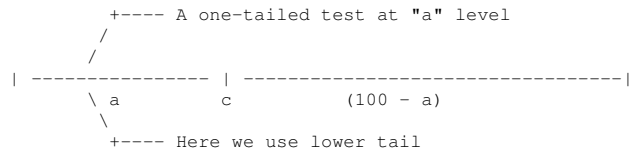
So we need to find the cut off point: if the number of cured falls within critical region, the we will say there is sufficient evidence to reject the null hypothesis. We will call this cut off point for critical region as **critical value**, c.

To find out critical region of the hypothesis test, we first decide on **significance level**, represented by α. It is a measure of how unlikely you want to result of the sample to be before you reject null hypothesis.

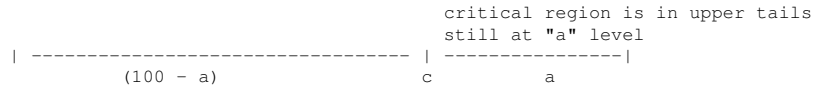
For this example: if we use X to represent the number of people cured, then we define critical region as being values such that:

$$P(X < c) < \alpha, \text{where } \alpha = 5\%$$

We also need to be aware that we are doing **one-tailed** test here: where critical region falls at one end of the possible set of values.



If your alternative hypothesis has $<$ sign, then use the lower tails, where the critical region is at the lower end of the data. If your alternative hyposis has $>$ sign, then use the upper taii, where the critical region is at the upper end of the data.

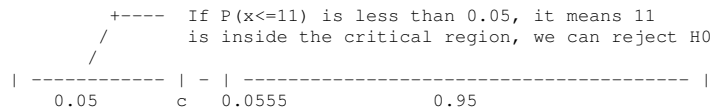


A **two-tailed** test is where the critical region is split over both ends of values. You still choose a level α, both ends contains $\alpha/2$, and total is α. You need two-tailed test when H_1 contains $\neq$ sign. In our example, $p \neq 0.9$ means you have to check significantly more people or significantly fewer people get cured.

Step 4: Find p-value

A p-value is the probability of getting a value up to and including one in your sample in the direction of critical region. This value depends on critical region and test statistic.

For this example, 11 people cured, and critical region is the lower tail of the distribution. This means the $p-$ value is $P(X \leq 11)$.



$$P(X \leq 11) = 1 - P(X \geq 12) = 1 - 0.9445 = 0.0555$$

Step 5: Is the sample result in critical region?

Our critical region is the lower tail of the probability distribution, and we use a signifance level of 5%. This means we can reject the null hypothesis if our p-value is less than 0.05. However, it is not.

Step 6: Make the decision

Since we can't reject the null hypothesis, we accept the drug company's claim.

10.2 Approximate with normal distribution

Say after a while, doctor collected more data, now his sample size is 100, with 80 cured, 20 not cured. We want to re-test the hypothesis. Now $X \sim B(100, 0.0)$.

Since n is large, and both $np > 5$ and $nq > 5$, we can use

$$X \sim N(np, npq), \text{or } X \sim N(90, 9)$$

as our test statistic.

If we standarize this, we get:

$$Z = \frac{X - 90}{\sqrt{9}} = \frac{X - 90}{3}$$

This means that for our test statistic, we can use Z, and $Z \sim N(0, 1)$.

There are two ways to come to the same conclusion:

Use test statistic value

Since the significance level is 5%, this means our critical value c is the value where

$$P(Z < c) = 0.05$$

Probability table tells you at probability 0.05, $c = -1.64$. In other words, $P(Z < -1.64) = 0.05$. So if our test statistic is less than -1.64, we have strong evidence to reject.

Test statistic $Z = (X - 90)/3$, 80 people were cured, so by substituting X with $x = 80$, $z = -3.33$, which is the value of our test statistic.

Use critical region

We can also use $z = -3.3$ to find out the probability of 80 or fewer being cured.

The p-value is given by $P(Z < z) = P(Z < -3.33)$. The probability table says $p = 0.0004$, less than 0.05.

Both methods come to the same conclusion: there are strong evidence to reject the null hypothesis.

10.3 Two Groups, Continuous Data

Example: Average fuel efficiency for 4-cylinder vehicles is greater than the average fuel efficiency for 6-cylinder vehicles. We define:

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 > \mu_2$$

Say, the two groups of cars we sampled are less than 30 in each. So we need to use $t-$ distribution. The test statistics is given by:

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{37}$$

where s_p is the pooled standard deviation:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \tag{38}$$

However, if the sample size is greater than 30, then regular normal distribution can be used for the sample distribution of mean difference, and the following test statistics can be used:

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{39}$$

10.4 Two Groups, Categorical Data

Assuming sample size is greater than 30.

$$H_0 : \pi_1 = \pi_2 \qquad H_1 : \pi_1 < \pi_2$$

Then the test statistics for sample distribution of difference of proportions is:

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{40}$$

10.5 Type I and II errors

A Type I error is what you get when you reject the null hypothesis when the null hypothesis is actually correct. It is like putting a prisoner on trial and finding him guilty when he's actually innocent.

The probability of getting Type I error is the probability of your result being in the critical region, which is defined by the significance level of the test, which is α. So,

$$P(\text{Type I error}) = \alpha$$

A Type II error is what you get when you accept the null hypothesis, and null hypothesis is actually wrong. It is like letting a guilty prisoner getting away from. Type II error is labelled as β.

How to find β

1. Check you have specific value for H_1
2. Find the range of value outside of critical region of the test
3. Find the probability of getting this range of values, assuming H_1 is true.

Still using previous example,

$$H_0 : p = 0.9 \qquad H_1 = 0.8$$

The values that are outside the critical regions are given by $Z \geq -1.64$. Since $Z = (X - 90)/3$, this means $X \geq 85.08$.

All we need to do is to work out $P(X \geq 85.08)$ assuming H_1 is true. And when H_1 is true, we can approximate probability distribution of X with $N(np, npq)$, where $n = 100$, $p = 0.8$, that is:

$$X \sim N(80, 16)$$

The standard score of 85.08 is:

$$z = \frac{85.08 - 80}{\sqrt{16}} = 5.08/4 = 1.27$$

In order to compute $P(X \geq 85.08)$, we can use probability table to find $P(Z \geq 1.27) = 1 - P(Z <\leq 1.27) = 1 - 0.8980 = 0.102$, which is easy enough. This is our β value for Type II error.

10.6 Power

The **power** of a hypothesis test is the probability that we will reject H_0 when H_0 is false. In other words, it is the probability that we will make the correct decision to reject H_0.

$$\text{Power} = 1 - \beta$$

11 Chi-squre Test

The χ^2 probability distribution has two key purposes:

1. It is used to test **goodness of fit**. This means that you can use it to test how well a given set of data fits a specified distribution.

2. Another use of the χ^2 distribution is to test the **independence** of two variables. It is a way of checking whether there is some sort of association.

11.1 Test Statistic of χ^2

To use χ^2 distribution to assess difference, we need a test statistic, it is defined as:

$$X^2 = \sum \frac{(O-E)^2}{E} \tag{41}$$

where O refers to the observed frequency and E refers to the expected frequency.

The χ^2 distribution takes one parameter, ν, the **degree of freedom**. It is the number of independent variables (or independent piece of information) used to calculate the test statistic X^2.

When ν has a value of 1 or 2, the shape of χ^2 distribution follow a smooth curve, starts off high and getting lower. The shape suggests that getting low value of the test statistic X^2 is much higher than getting high values. In other words, observed frequencies are likely to be close to the frequency you expect.

When ν is greater than 2, the shape of χ^2 distribution changes: it starts of low, gets larger, then decreases again as X^2 increases. The shape is positively skewed, but when ν is large, it is approximately normal.

A shorthand way of saying that you are using the test statistics X^2 with χ^2 distribution that has a particular value ν is:

$$X^2 \sim \chi^2(\nu)$$

11.2 Hypothesis Testing Using χ^2 Distribution

Say, we have the expectation and observation table such as:

x	-2	23	48	73	98
$P(X=x)$	0.977	0.008	0.008	0.006	0.001
Frequency	965	10	9	9	7

Decide significance and critical region

When you conduct a test using χ^2 distribution, you conduct a one-tailed test using the upper tail of the distribution as your critical region. This way, you can specify the likelihood of your results coming from the distribution you expect by checking whether the test statistic (X^2) lies in the critical region of the upper tail. If you conduct a test at significance level α, then you write this as:

$$\chi^2_\alpha(\nu)$$

To use χ^2 probability table and find critical value, you need degree of freedom, ν and α. The value given by the intersection is x, where $P(\chi^2_\alpha(\nu) \geq x) = \alpha$. For example, if $\nu = 8, \alpha = 0.05$, then you locate $x = 15.51$.

That means if test statistic X^2 was greater than 15.51, then it is in critical region at 5% significance level with 8 degree of freedom.

For above example: $\chi^2_{5\%}(4) = 9.49$ and $X^2 = 38.272$ based on Eq. 41. Therefore we reject the null hypothesis that the observed frequency match the expected value.

This type of hypothesis test is also known as **goodness of fit** test. It tests whether observed frequencies actually fit in with an assumed probability distribution. It works for pretty much any probability distribution. The key is to find ν, the degree of freedom.

ν for different distributions

- Binomial: if you know what p is, then $\nu = n-1$; If you don't know p and you have to estimate it from observed, $\nu = n-2$.
- Poisson: You what what λ is, then $\nu = n-1$; If you don't know *lambda* and you have estimate from observed, $\nu = n-2$.
- Normal: You know what μ and σ^2 are, then $\nu = n-1$; You don't know what they are and have to estimate from observed, then $nu = n-3$.

11.3 Independence Test Using χ^2 Distribution

An *chi*^{2} test for independence is a test to see whether two factors are independent, or whether there seems to be some sort of association between them. In the casino example: we want to test if the croupier leading the game has any impact on the outcome. In other words, we assume the choice of croupier is independent of the outcome, unless there are strong evidence against it.

When you are just given a table like this:

	Croupier A	Crouper B	Croupier C
Win	43	49	22
Draw	8	2	5
Lose	47	44	30

You have to calculate their expected frequency first, by building so called **Contingency Table**. Basically, it adds a total row at the bottom and a total row at the right-most column and a grand total. Then, expected frequency is calculated as:

$$\text{Expected Frequency} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}} \tag{42}$$

Once expected frequencies are known, you can use Eq. 41 to obtain test statistics X^2.

Now another crucial thing you have to know: degree of freedom. Given a table with h rows, k columns, the degree of freedom is:

$$\nu = (h-1) \times (k-1) \tag{43}$$

The key observation is: if you fix on one row, and just look the corresponding column. You only need to know $k-1$ of the columns as the total frequency of the row is fixed and known, thus the constraint.

The rest of the hypothesis test process is the same as before.

12 Correlation and Regression

- Univariate data deals with just one variable. Bivariate data deals with multiple variable.
- A scatter diagram shows you patterns in bivariate data.
- Correlations are mathematical relationships between variables. It doesn't mean that one variable causes the other. A linear correlation is one that follows a straight line - a linear correlation can be positive (upward trend) or negative (downward trend)
- The line that best fits the data points is called **the line of best fit**.
- Linear regression is a mathematical way of find the line of best fit, $y = a + bx$.
- The sum of squared errors, or SSE, is given by $\sum(y_i - \hat{y}_i)^2$. y_i represents y value in the data set; $\hat{y}_i$ represents estimate using line of best fit.
- One mathematical way of finding the line of best fit (or a and b) is known as **least square regression**.

  The slope of the line $y = a + bx$ is

  $$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

  The value of a is given by

  $$a = \bar{y} - b\bar{x}$$

  The line given by $y = a + bx$ is called the **regression line**.
- To describe the scatter of data away from the line of best fit, we use the **correlation coefficient**, r, which is a number between -1 and 1.

  if $r = -1$, there is a perfect negative linear correlation; if $r = 0$, there is no correlation; if $r = 1$, there is perfect positive correlation. You find r by

  $$r = \frac{bs_x}{s_y}$$

  s_x is the standard deviation of the x values in the sample.

  $$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Similarly, s_y is the standard deviation of the y values in the sample.

$$s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}$$

- An **alternative notation for least square regression** is to define **covariance**, s_{xy}. Just as variance s_x describes how x values vary, and s_y describes how y values vary, the covariance of x and y is a measure of how x and y vary together.

$$s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

Then, we can re-write formula for b as:

$$b = \frac{s_{xy}}{s_x^2}$$

And re-write formula for r as

$$r = \frac{bs_x}{s_y} = \frac{s_{xy}}{s_x s_y}$$